32 | 服务的容量规划：怎样才能做到有备无患？

2020-02-07 庄振运

讲述：庄振运
时长 16:17　大小 13.05M

你好，我是庄振运。

今天我们来讨论一下在公司运营方面很重要的**容量规划**。容量规划，就是根据互联网服务的需求和公司发展目标，而决定容量的供应能力的过程。

光说概念你可能不太明白，不过你可以这么理解，容量规划是为了回答一系列和公司业务运营有关的重要问题而产生的：

- 单台服务器的最大处理能力是多少？
- 未来半年公司还会有多少可用容量？

春节前如果要进行一次大促，当前的容量能不能支撑？

……

今天我先讲容量规划的目标和基本过程，然后再一个一个地讲容量规划的六大挑战。

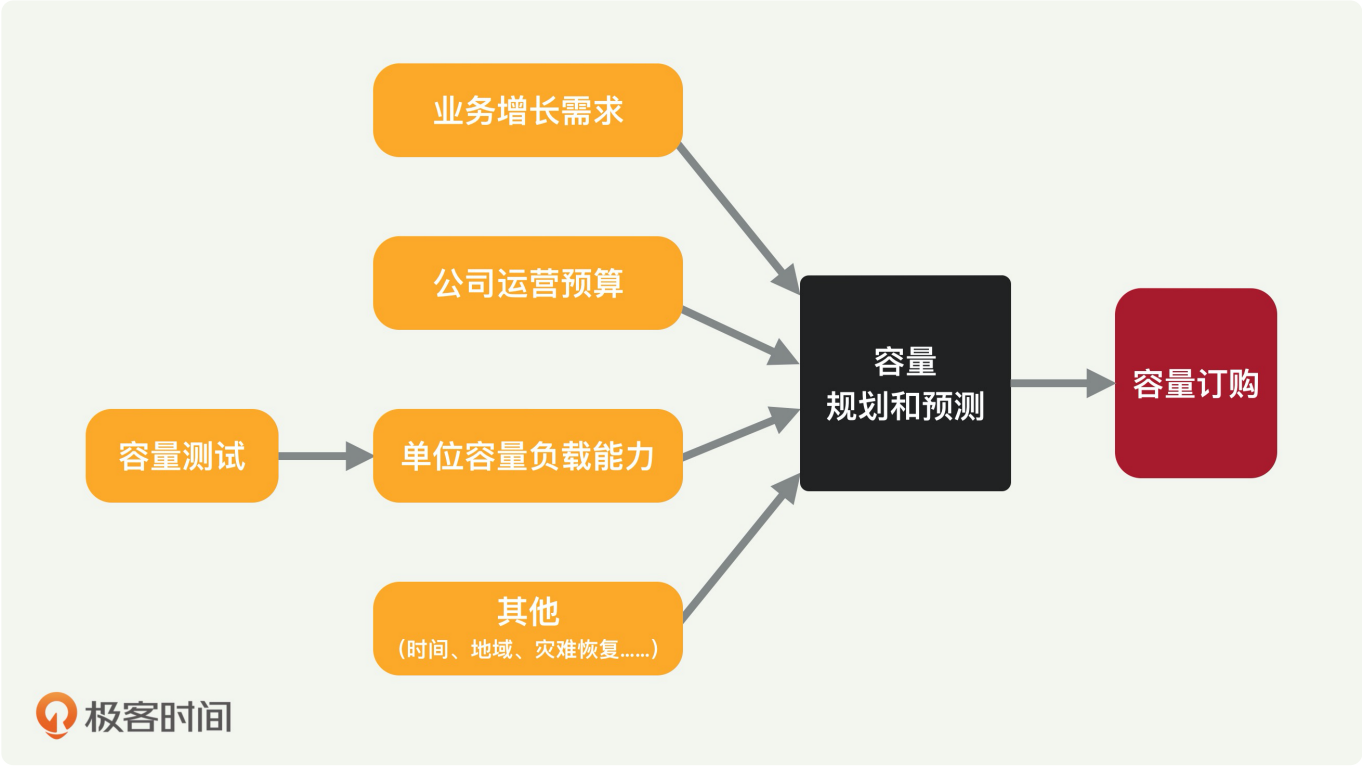
我们前面也大概提到过，容量规划的目的有两方面。一方面是为了保证公司的正常运营和业务增长，及时地提供足够的容量，来满足未来所需。另一方面，也是希望空闲的容量越少越好，因为每一台空闲的服务器，都消耗公司的运营成本。

现代的互联网服务规模的扩展，主要是用“水平扩展”而不是“垂直扩展”（通过增加服务器的数量和网络的带宽来提供更多的容量，而不是通过升级服务器）。相对于垂直扩展，水平扩展的优点显而易见，就是成本相对较低、调整容易、扩展性好。

怎么做容量规划？

整个容量规划的过程分为几部分：首先是容量测试来决定单位容量的负载能力；同时确定公司的业务增长需求，并且获取公司的运营预算。然后集成其他的考虑因素（包括时间、地域、灾难恢复等），做出合理的规划和决策。最后根据决策结果，进行容量订购。

为了方便把握，我把这几个因素模块整合在一起，如下图所示，我们一个模块一个模块地讲。



容量测试决定单位容量的能力

容量测试可以决定单位容量的能力。容量规划的基础，就是确定单位容量，比如一台服务器系统的处理能力。而进行容量测试，一般是让服务器处于最大负载状态，或某项性能指标达到所能接受的最大阈值下，对请求的最大处理能力。

容量测试是性能测试里的一种测试方法，它的目的就是**测量系统的最大容量**，为系统扩容、性能优化提供参考，从而达到节省成本投入，提高资源利用率的目的。

容量测试大致分为两种：线上测试和线下测试。我们在第 25 讲专门讲了线上测试，这里快速地复习一下。线上测试相对比较准确，因为是用真实的生产流量负载。线下测试，则比较容易进行。但是你要特别注意测试过程和结果的代表性，就是它能否真实客观地反映生产环境中的数据。尤其是测试环境的配置和驱动数据，一定要和线上的生产环境尽量保持一致。

预测未来需要的容量需求

要做好容量规划，除了单位容量的处理能力，另外一个必须要知道的输入，是服务和业务的需求。通俗点说，就是期望达到的用户数和数据流量大小。比如一个前端服务，要决定明年的容量需求，就需要确定明年的预期用户数目。

对于未来的容量需求的预测，主要是参考历史的相关数据，然后建立合适的模型。比如我们在第 26 讲提到的针对时间序列的 ARIMA 模型，就是一个不错的模型。如果针对的业务已经有足够长的历史，那么过去的数据就很值得参考。

反之，如果一个业务还没有上线，或者上线时间很短，那么历史数据就不是很可靠。如果你面对的是这种情况，有两种解决方法：一是可以类比其他相似的服务，不管是公司内部的还是业界的，都可以拿来参考。二是需要借助一些主观的推测。当然这种主观推测也不能凭空乱猜，而是要有理有据。

容量规划时，公司的业务运营预算也要考虑进去。一个正常运行的公司，运营成本总是有限的（每个业务都有预算定额）。如果对一个业务预期的容量需求，超过已经设定的预算定额，那么，你要么是把容量需求降低到预算定额，要么就得向公司申请增加预算。

考虑各种限制因素和挑战

另外还有一点要注意的是，未来容量的需求预测总会和实际情况有差距，也就是说，总是有不确定的因素。

所以，我们一般会根据估计的信心指数，对预测结果做几种不同的预测。比如 90% 的信心和 50% 的信心。这里信心指数越大，容量数值也越大，就是说越有把握满足需求。举个例子，比如我们要估计明年一个服务的容量需求，可能得出结论：

如果只需要保证 50% 的信心，1000 台服务器就够了；

要保证 90% 的信心，则需要 1200 台服务器。

但是，无论我们多么努力地去预测，总有些因素没有考虑到，从而导致实际情况和预测不吻合。所以，我们一般都要对容量的需求预测加上一个冗余 Buffer，我个人的经验是 10% 的 Buffer。比如我们的模型预计需要 1000 台服务器，那就加上 100 台的 Buffer，也就是需要 1100 台服务器。

及时和实时地调整

也正因为容量的规划和使用过程中，有很多的不确定因素，随着时间的推移，这些不确定因素慢慢就会变成确定的因素。比如，在一个服务还没有上线的时候，做容量规划非常困难，因为不能准确地预测用户数量。但是随着服务的上线，开始积累用户量，对未来用户数量的预测就会变得越来越容易。

所以，容量的预测和规划，也是一个持续和不断迭代的过程。开始的时候靠模型，靠历史数据和一些直觉来做大体预测。然后，定期地和实际的数据做对比，从而调整和纠正以前的预测。

容量规划的挑战在哪?

容量规划和预测时，除了前面讲的步骤和因素外，还有其他几个限制因素和挑战值得考虑。我总结了六个这样的挑战，包括地域的限制、服务器种类的区分、时间和季节性等等。

容量规划不仅仅是钱的问题

第一个挑战就是，容量规划不仅仅是钱的问题。你可能听说过：“能用钱解决的问题，都不是问题。”

但容量的供应问题，很多时候还真不是砸钱就能解决。

为什么有时候钱不能解决容量问题呢？如果公司的运营规模很大的话，需要的容量供应也就很大。在很多情况下，市场上能够找到的数据中心很有限，所以即使公司愿意投入很多资金，也不一定就能在需要的地点和需要的时间租赁到足够的容量。

有的公司自己建造数据中心，那就面临更多新的的问题和挑战，尤其是时间方面。建立数据中心的交货时间至少是两、三年。数据中心建造和扩建，需要寻找可用的建筑商和建筑工人，但是找到足够的合格建筑商和工人并非易事。因此，数据中心的建设并不像把钱扔在问题上那么简单，需要提前很久就规划和考虑，要未雨绸缪。

数据中心的地域因素

第二个挑战是地域因素。当公司的业务需要部署服务器在多个数据中心时，每个数据中心都需要做相应的容量规划。也就是说，仅仅只有预测和规划一个总体的容量是不够的，还需要具体到每个数据中心的内部。

数据中心往往分布在不同的地域，它们内部的服务器，并不能在数据中心之间随意置换。为什么呢？因为我们的业务也要求服务器尽量地离客户近些，从而降低网络延迟，提高带宽和可靠性。像 Facebook、Google、腾讯、阿里这样的公司尤其是如此，它们往往在全球有很多数据中心，而且用户的规模很大，如果容量规划做不到数据中心这个层次，是肯定不行的。

服务器种类的因素

下一个挑战是服务器种类的挑战。公司往往会有不同种类的服务器，以便针对特定类型的工作负载进行量身定制。例如，有专门针对内存密集型工作负载的大内存服务器。当有多种服务器可选的时候，容量规划就需要选择最合适、最经济的服务器类型，这样可以降低成本并改善服务质量。

对于一种具体的服务器而言，它的性能也是随着时间的推进不断改善的。比如一种计算密集型的服务器，虽然基本的机架设计不会改变，但是服务器使用的 CPU 还是会一直更新换代的，一般每年都会有新的 CPU 型号。

所以，在做容量规划时，也要把这个因素考虑在内。比如要支撑一个服务，如果用现在的 CPU 型号预测，明年需要一千台服务器。可是同时明年的 CPU 会升级，性能提高了 10%。那么考虑到这点，可能只要 900 台服务器就够了。

时间和季节的因素

容量规划也要充分考虑时间和季节的因素。在第 26 讲里，我们一起看过正常生产环境的负载变化。对大多数互联网服务来讲，在一天内，每天的上午上班时间的负载较大；在一周内，工作日的业务负载比较大。同时，当经历节日时，比如春节、国庆等，也要注意流量的预期变化。

对于面向普通用户的社交网站，比如 Facebook，每当新年、圣诞节时候，有些业务的用户使用量会好几倍的增长。这些业务包括照片和视频上传，朋友分享等。所以我们的容量规划过程会充分考虑到这些时间的业务需求，提前部署足够的容量来迎接这些峰值的服务流量。

不同的灾难恢复场景

还有，为了保证业务的可靠性，灾难恢复（DR，Disaster Recover）场景是必须考虑的。

灾难恢复准备，意味着公司的容量可以容忍某个容量单位的损失。互联网服务其实有很多种不同的 DR 场景，最常见的是数据中心的 DR，也就是说，假如一个数据中心因为天灾人祸的原因完全不能访问，其余的容量如何支撑整个互联网服务的正常运作。

灾难怎样影响服务性能呢？损失的容量单位，会导致其他容量单位的负载增加。负载的增加量，取决于丢失单位在整个容量系统中的比例。具体来说，假设丢失的容量单位，平时承担的负载百分比为 x%，那么在丢失该容量单位之后，所有其他单位的流量将增加[1 /（1-x%）-1] * 100%。

为了帮助你理解，我先举个简单例子。假设一个服务有两个数据中心支持，这两个数据中心大小是同样的，都分别承担 50% 的服务流量。那么，如果一个数据中心不能用了，另外的数据中心就需要承担 2 倍的服务流量，也就是 1/（1-50%）=2。

再举个稍微复杂的例子。考虑以下情形，其中 A、B、C 的三个容量单位分别占流量的 40%、40%和 20%。如果单位 A 丢失，那么单位 B 和 C 的流量将增加 1 /（1-0.4）-1 = 67%。如果 C 单位丢失，则 A 和 B 单位的流量将增加 1 /（1-0.2）-1 = 25%。

不同服务之间的互相影响

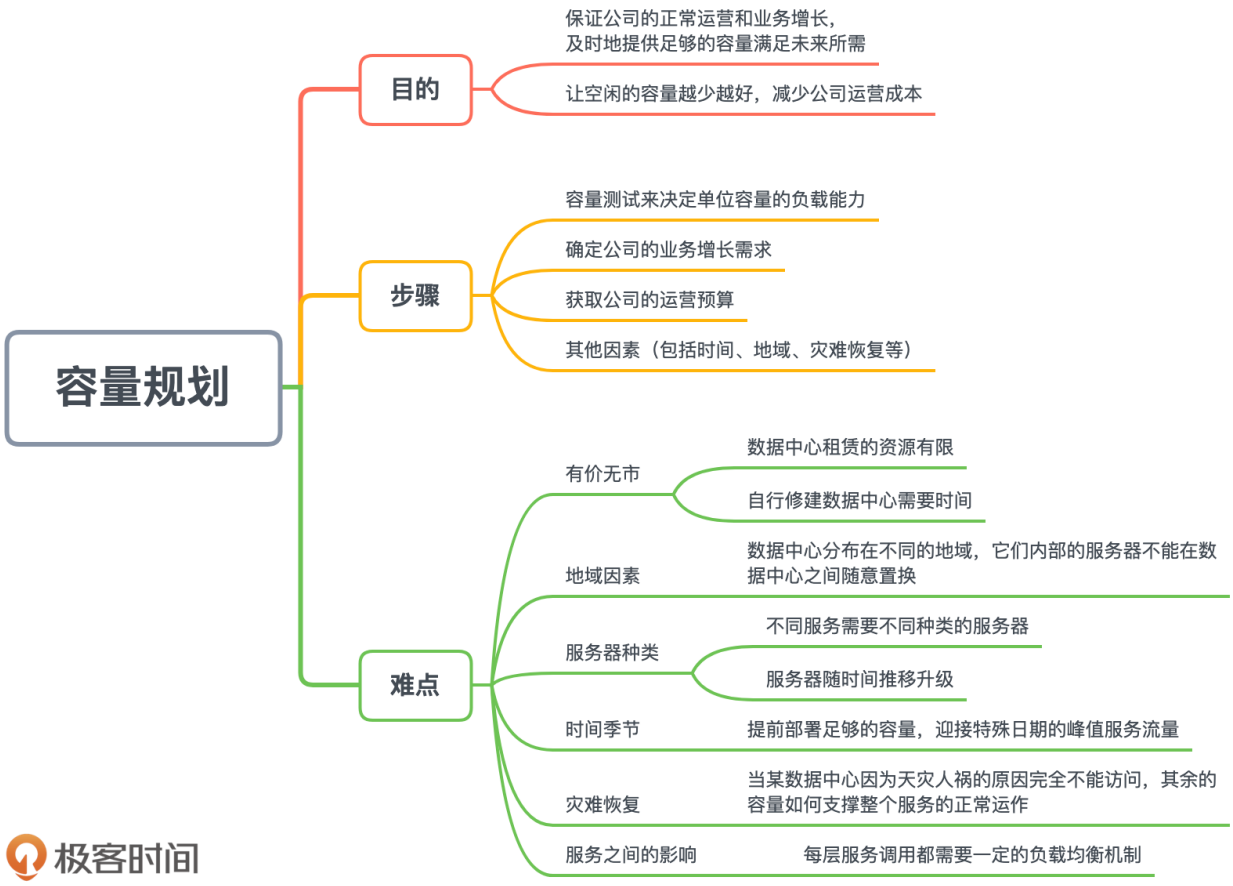
最后，互联网公司的业务，往往会分解成很多的服务。面对外部客户的是前端服务；还有很多后端和内部的服务来支撑前端服务。所以，如果前端服务的负载有变化，后端服务也会受到影响。要预测后端服务的容量需求，也必须充分考虑前端服务，以及其他上下游服务的影响。

一般来讲，每层服务调用都会有一定的负载均衡机制，在规划每层服务的容量需求，以及每层服务的不同容量单元的容量需求时，就要了解这些负载均衡机制，从而确定相关因子。比如，如果前端服务负载增加 100%，则下游服务预计会增加多少负载？如果某个下游服务的负载增加也是 100%，则相关因子为 1。

每个服务的总体负载确定后，还要考虑具体到这个服务在不同数据中心的分配，也就是需要确定每个数据中心分别分配多少。

总结

我们今天讨论了容量规划的过程和需要注意的地方。容量规划这个工作非常重要，正如清朝人陈澹然讲的：“不谋万世者，不足谋一时；不谋全局者，不足谋一域”。如果规划做不好，公司业务一定会受到严重的影响。要不就是没有容量可用，要不就是容量太多造成浪费。



容量规划的工作并不好做，你从我说的几大挑战就看得出。所以需要对公司的业务特点和服务架构非常熟悉，并不断及时地调整，才能尽可能地达到预定的目标。

思考题

你们公司或者你的部门的业务，需要什么样的容量来支持呢？比如何种服务器？

对于容量规划，你们部门对未来半年甚至一年的容量要求是多少？是哪个部门帮你们做容量规划，他们是怎样确保容量规划的准确性呢？

欢迎你在留言区分享自己的思考，与我和其他同学一起讨论，也欢迎你把文章分享给自己的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　31 | 规划部署数据中心要考虑哪些重要因素?

下一篇　33 | 服务效率提升：如何降低公司运营成本?

精选留言 (1)

写留言

黄海峰
2020-02-07

我们就是程序员自己容量规划，而且领导和其他上游都觉得这事很简单，现在说了要求很快就要给出结论。。。所以很多时候我们也是随便看几个指标然后拍脑袋给个结论

作者回复: 做一个工作做久了，就会慢慢特别有感觉。哪怕是拍脑袋一下，经常给出的数字也会八九不离十。恭喜你:)